建设微生物组大数据中心发挥长期科学影响*



张国庆[1**] 宁 康[2] 职晓阳[3] 刘 婉[4] 徐 萍[5] 周豪魁[6] 胡黔楠[1] 赵国屏[1**]

1 中国科学院上海生命科学研究院生物医学大数据中心 上海 200031
2 华中科技大学生命科学与技术学院 武汉 430074
3 云南大学生命科学学院微生物研究所 昆明 650091
4 上海生物信息技术研究中心 上海 201203
5 中国科学院上海生命科学研究院生命科学信息中心 上海 200031
6 中国科学院深圳先进技术研究院合成生物学工程研究中心 深圳 518055

摘要 宏基因组研究的思想与技术推动了微生物组的兴起，积累了丰富的微生物基因组以及健康、动植物和环境相关的微生物宏基因组数据，形成了具备一定规模和影响力的数据库、标准化方法与分析工具。大多数平台聚焦于为项目或特定类型的微生物菌群提供数据支撑，难以满足更深入全面的微生物生物学研究需求。文章建议采用综合聚焦微生物分类单元总和的微生物系统组与聚焦特定生态位微生物种群总和的微生物组的思路，建设综合性的微生物组数据仓库，整合微生物分类、进化、生态以及相关“组学”数据与信息。在此基础上，进一步综合生命科学基础研究和系统合成生物学研究的数据，支撑经高水平质控的综合性参考数据库、标准化的拼接与注释以及一流的数据汇交、搜索分享、深度学习和分析挖掘方法的研究开发。由此，亦将进一步集成大型微生物组项目的元数据及数据，形成数据综合完整、管理安全高效，服务功能完备的微生物组大数据中心。

关键词 微生物组，微生物系统组，分类，生态，合成生物学

DOI 10.16418/j.issn.1000-3045.2017.03.009

*资助项目：国家高技术发展“863”项目计划（2014AA021502、2015AA020108），国家重点研发计划（2016YFC0901904、2016YFC0901604）

**通讯作者

修改稿收到日期：2017年3月1日

自从2005年10月成立国际宏基因组联盟以来，多个国家启动了人类微生物组相关的研究计划。包括美国的人类微生物组计划HMP[1]以及后续项目iHMP[2]，欧盟的MetaHIT[3]以及后续项目MetaGenoPolis，标准项目IHMS[4]，韩国也启动了微生物组多样性项目。此外，国际上开展了大量以人类健康疾病及环境检测、修复为主要目的科研相关的微生物组项目。

这些项目的实施，推动了专业数据库和参考数据库、标准化与质量控制、数据分析挖掘工具等平台性支撑工作的发展。

当前的微生物组项目公开产出的数据大多属于基础数据，可以通过EMG[5]、NCBI、JGI等第三方的数据中心发布，例如TaraOceans[6]、MetaHIT[3]、HMP[1]、GOLD[7]等项目的数据，相关数据资源发布情况见表1。有些数据资源平台，除了提供数据访问功能外，还提供在线分析注释功能，例如JGI IMG/M[8]、MG-RAST[9]和iMicrobe（http://imicrobe.us/）等平台。

MIxS是GSC制定的序列最简描述信息标准集，其中MIMS是MIGS的延伸标准[10,11]。MIMS标准为不同的环境制定了通用的“环境包”供各个项目共用，其中包括空气、建筑内环境、人类相关、人类口腔、人类肠道、人类皮肤等15个环境包，已经成为NCBI、MG-RAST[9]和GOLD[7]等主流数据库的样本描述指南。M2B3[12]被用于海洋生物多样性相关的分子生物学样品元基因组测序项目的标准规范。

当前国际上的微生物组数据平台主要围绕参考数据目录和元基因组数据，为某种或某类生态环境的微生物组研究（项目导向为主）提供功能注释、群落物种结构解析等基础性分析服务，因此，只能说是某一环境的微生物“分子生态学”与元基因组的研究平台。对微生物学其他研究，如微生物系统分类、综合性生态分析、元件库构建和细胞工厂设计，只能提供数据下载之类的低层次支撑，难以提供更直接的在线计算甚至算法测试环境的支撑。此外，部分平台的数据仅覆盖特定项目，数据覆盖度不够充分，整合度不够综合，需要研究人员花费大量时间来发现、获取、整合数据资源、信息资源和知识资源，难以实现简便易行的目标。

因此，我们认为，现在通用的“微生物组”的概念，即“微生物组”（microbiome）是存在于特定环境（生态位，biotype）里的多种类微生物群（microbiota）的所有成员及其遗传信息（主要是meta/megagenome）和生命功能的集合；需要与一个更宽泛的“微生物系统组”（microbiophylome）概念相联系。它是“所有”微

表1　常见的微生物组数据库*

数据库	EMG	IMG/M	GOLD	iMicrobe	MG-RAST	NCBI
建库时间	2013	2005	2015	-	2008	-
建库单位	EBI	加利福尼亚大学	加利福尼亚大学	亚利桑那大学	芝加哥大学	NCBI
国家	英国	美国	美国	美国	美国	美国
项目量	>1 000	258 (公开)	>1 000	261	-	>15 000
样本量	>60 000	3 515（公开）	>20 000	5 171	-	>400 000
网站	https://www.ebi.ac.uk/metagenomics/	https://img.jgi.doe.gov	http://www.genomesonline.org	http://imicrobe.us/	http://metagenomics.anl.gov	-
数据访问在网站	否	否	否	是	是	否
数据访问网址	ENA网站	JGI网站	JGI网站	本数据库网站	本数据库网站	NCBI
数据下载方式	Aspera、网站页面	网站UI、Globus和JGI API	网站UI、Globus和JGI API	网站页面、API、iPlant Discovery环境、CyVerse数据镜像服务、FTP	网站页面、FTP	SRA下载工具
数据提交	是	是	是	是	是	是
开放程度	部分公开	部分公开	部分公开	公开	部分公开	公开
是否提供在线分析	是	是	是	是	是	否
提供可视化工具	是	是	是	是	是	否

* 截至2016年12月

生物个体（microbe）以及各种微生物群体（microbiota）成员的遗传信息（主要是多种“组学/Omics”信息）及相关生物学结构功能的集合。整合这两个概念下的微生物组数据，建设具有全局搜索功能的数据仓库，通过不断发展的高质量拼接注释技术与研究服务平台，形成高标准的质量控制；进而实现高通量数据安全有效的汇交共享与分析挖掘。

上述“微生物组数据仓库”应该涵盖微生物分类、进化、生态三类数据，作为这三大类数据的管理共享平台，是为建立高质量的数据标准和数据质控流程开展长期发展的研究的基石；而在提供高质量数据标准和数据质控流程的基础之上，数据仓库又将进一步全面收集国际大型项目产出数据的元数据，并按需集成原始数据，形成数据完整、功能完备的微生物组大数据中心。这个大数据中心是包括微生物组研究项目的所有微生物学研究数据的基础性平台，将提供微生物组系统分类学工具包、微生物组系统生态学工具包、微生物组系统合成生物学工具包等挖掘开发的工具系统（图 1）。

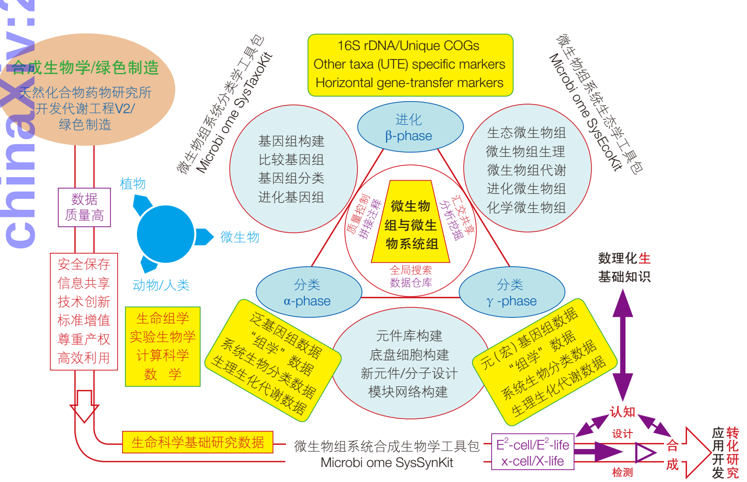


图 1　微生物组大数据中心建议实施图

1 微生物组系统分类学

微生物系统分类是微生物组研究的理论基础之一，经历了形态分类、化学分类、分子分类 3 个阶段，最终形成了多相分类这一技术与理论体系[13]。16S rRNA 系统发育提供了现行分类系统的基础框架，但是在较低级分类单元上的低分辨率一直受到学界的诟病。新一代高通量测序技术正在深刻地影响着微生物系统分类学的发展，在对新物种的多相分类鉴定中整合基因组数据是分类学发展的趋势[14]。微生物基因组高度的结构与基因多样性以及横向基因转移使得对进化历史的解析问题变得更加困难和复杂[15]。将基因组纳入多相分类体系仍需要一个过程[16]。

16S rRNA 之类的标记分子已经发展了较为完善的大型特征序列数据库 ,如 RDP[17]、Greengenes[18]、Silva[19] 等。但是由于自然界中存在着为数众多的未培养微生物，大量测序数据仍无法确定分类地位，要了解目标微生物的功能仍然需要依托纯培养微生物基因组数据作为参考。一定程度上，微生物组研究理论意义上的瓶颈是微生物分离纯培养技术与完善的分类系统。大量的微生物组研究项目产出的数据构成了进行元分析的参考数据库，给研究者提供了背景信息，有助于通过元分析找到新结论。

16S rRNA 基因之所以成为现代分子分类的黄金分子，除了功能保守性以及适度的进化速率外，更重要的是几十年积累的数据库资源，目前大部分已鉴定的微生物物种都有对应的基因序列信息[17-20]。基于基因组数据在分类学中的应用同样需要一个数据平台，便于数据的存储和分析。国际上已有的基因组数据子库包含大量冗余数据，不利于基因组的比较分析，更缺乏分类学数据支撑。因此，急需一个整合的数据库平台，在分类学上有效描述物种基因组和分类鉴定表型数据，并提供数据分析流程。通过统一的优化标准和分析流程对每个基因组序列进行基因预测、功能注释、代谢网络重建，支持深入挖掘基因组信息；整合分类表型数据便于研究者查询、比较分析不同物种的特征，开展表型的遗传机制等比较基因组学研究工作。

2 微生物组系统生态学

元基因组是普遍用于目前不可培养的微生物研究的

菌群结构与功能的代表性方法，全长 DNA 测序和 16S rDNA 是典型的两种技术手段。这两种技术手段的结合，使得我们对一些重要微生物群落的结构和功能的认识迅速取得了突破[21,22]。面向这两种方法产出的数据，可以开展质量控制、序列归类与功能划分、集成方法等生物信息分析（表 2）。

（1）元基因组序列质量控制涉及到一系列的短序列质量分析与过滤。最重要的序列质量控制步骤包括：短序列质量分析，短序列修剪，嵌合体短序列的去除等。短序列质量控制可以通过 Mothur[23]和 QC-Chain[24,25]等软件包来实现。

（2）根据测序数据的质量差异和序列长度，测序数据可以被归类到“门”“纲”“目”“科”“属”中不同精确度的层次。序列归类（classification）可以被分为序列比对（similarity-based）分析和序列成分（composition-based）分析。序列比对分析受限于已知归类和功能的序列，90% 以上的微生物群落测序数据无法通过这种办法进行归类。序列成分分析的方法依赖于序列 GC 含量、编码区比例等特定特征，通过和已知基因组中的特征向量比较，确定序列归类[26]。MEGAN[27]借助 NCBI 分类数据库，展示了单一或者多个元基因组中各组分的进化位置和分类组成。

（3）元基因组的研究对象可以分为群落物种结构和群落功能结构两方面。近年来，以 16S rRNA 生物标记为基础的分析技术，例如 MOTHUR[23]、QIIME[28]、Parallel-META[29]等，拓展了微生物群落结构的研究范围，但其高保守性和多拷贝性也使其应用范围受到限制。在功能结构上，元基因组学的基本研究策略包括大片段 DNA 的拼接、基因预测、基因注释以及代谢通路分析等。此外，考虑微生物不同群落的特点（基于群落元数据），可以将所有数据分为两个或以上的组（class），进而开展群落生物标记的识别和鉴定，开展基于微生物群落全基因组测序数据的群落功能特征标记挖掘。

（4）基于单一类型的数据的挖掘越来越无法满足微生物群落研究的需求。代谢组、单细胞数据也逐步与元基因组数据整合。在群落代谢物组方面，小分子代谢物、核磁共振标准谱图、标准质谱谱图，以及各代谢物的相关物化信息，相关数据分析方法也正在发展过程

表 2　代表性的生物信息学分析平台

软件（平台）	数据库	分析数据对象	分析策略	分析结果
MEGAN	NCBI	16S rRNA	序列比对分析	物种结构，丰度和功能分类，以及物种之间的比较
ConStrains	整合数据库	宏基因组	序列比对和序列成分分析	物种结构，丰度
MetaPhlAn	整合数据库	宏基因组	序列比对和序列成分分析	物种结构，丰度
PICRUSt	整合数据库	宏基因组，16S rRNA	序列比对和序列成分分析	物种结构和功能分类
antiSMASH	整合数据库	宏基因组	序列比对和序列成分分析	BGC分析
CARMA	Pfam	16S rRNA	序列比对分析	物种结构和功能分类
Sort-ITEMS	NCBI	16S rRNA	序列比对分析	物种结构和功能分类
Phyloshop	Greengenes	全基因组，16S rRNA	序列比对分析	物种结构和功能分类
UniFrac	NCBI	16S rRNA	序列比对分析	物种结构，丰度和功能分类，以及物种之间的比较
QIIME		16S rRNA	序列比对分析	物种结构，丰度和功能分类
PhyloPythia	NCBI	16S rRNA	序列成分分析	物种结构和功能分类
MG-RAST	整合数据库	全基因组，16S rRNA	序列比对和序列成分分析	物种结构，丰度和功能分类，以及物种之间的比较
CAMERA	整合数据库	全基因组，16S rRNA	序列比对和序列成分分析	物种结构，丰度和功能分类，以及物种之间的比较
Galaxy	整合数据库	全基因组，16S rRNA	序列比对和序列成分分析	物种结构，丰度和功能分类，以及物种之间的比较

中。在单细胞数据分析方面，主要包括单细胞基因组、转录组和表征信号。以上数据分别从全局和个体、遗传和表观、结构和功能等不同角度为微生物群落研究提供支持。

从以元基因组为代表的微生物系统生态学相关的工具研究来看，除了工具自身的算法和性能外，工具背后的数据集的范围和质量会严重影响工具的准确性。构建统一的数据仓库，整理完整的特征序列、参考基因组、功能基因组等微生物系统组，以及典型微生物生态群落元基因组与代谢组等数据集，形成立体的完整的微生物组数据矩阵，开展元基因组拼接、注释、群体相互作用与网络、微生物生态比较等研究，形成微生物组系统生态学工具，发展生态微生物组、微生物组生理与代谢、进化微生物组、化学微生物组等特色微生物生态数据平台。

3 微生物组系统合成生物学

通过高通量测序的元基因组数据挖掘天然产物合成基因，主要通过三种方式：（1）单类化合物的基因簇注释，如 PRISM[30]和 GRAPE[31]。（2）单物种的基因簇注释，如 StreptomeDB[32]。（3）多物种多化合物基因簇注释，如 antiSMASH[33]。基因尺度的基因簇功能注释，只占基因数据的一部分。基因组尺度的挖掘不仅能够发现新颖的天然产物，而且还能发现相关合成途径，为生物合成研究提供了数据基础。

多个研究小组开发了不同的微生物细胞工厂相关的分子结构生物转化以及催化元件数据库，包括生物合成反应和催化元件数据库 BRENDA[34]，分子结构转化数据库 Rhea[35]，KEGG 分子结构转化数据[36]等。Rxnfinder 研究小组基于文献，开发了数据驱动型一站式生物合成新反应、新酶、新途径设计技术体系[37]。这些数据库从不同层次和角度包含了丰富的合成生物学资源， 但是以单个微生物为研究单位的生物学数据库还未全面建立。

为了能够高效并合理地开发设计目标化合物的生物合成路径，已经发展了多个基于原子匹配，或者反应规则的路径设计方法等[38-41]，但是极少以底盘细胞为研究单位。微生物细胞工厂的设计成为生物合成领域研究的重点[42]，产生了多个基于底盘细胞的目标化合物合成路径设计的方法，如 FMM[43]、PHT[44]、MRE[45]等。随着基因组测序技术的不断发展，以基因组尺度矢量代谢网络模型和基于通量平衡分析方法（FBA）[46]的模型优化方法为基础，实现合成目标化合物路径的设计，优化以及产量评估是目前模拟细胞工厂设计的另一个重要研究方向。

从合成生物学的发展进程来看，基因组尺度的合成设计正在越来越重要。依托微生物组的元基因组与功能基因组、转录组、代谢组等数据平台，建立合成生物学的微生物资源库，集成天然产物、调控催化与转运等元件、反应通路与网络、基因线路与基因簇等数据。研究底盘细胞及其代谢模型等全基因组尺度上的分析工具，发现新的贯通海量高噪音的微生物组数据与高易用性的基因簇和基因线路，形成微生物组系统合成生物学工具包。

4 微生物健康大数据应用

元基因组研究手段已经渗透到环境生物监测与治理[47-49]和极端环境[50]、营养与健康等以利用或克服复杂微生物群落及其产物为目的的科学领域。在医学领域，了解人体微生物群落结构与功能的变化有助于把握人类相关健康动态，尤其是在人体口腔环境[51,52]、肠道及其消化机制[53]、皮肤敏感度[54]等方面。在生物能源领域，复杂的生物能源过程如纤维素乙醇的转化与发酵[55]、沼气的生成[56]等，都是依赖于微生物群落的作用而完成。

在开展健康、环境、营养等方面的微生物组应用时，微生物组大数据中心以中立的第三方服务平台的方式，不仅能够为研究和应用提供数据资源、分析挖掘方式、知识库等方面的支持，而且能够形成公共数据与公共方法→私有数据在线分析并保存在公共平台→择时

与潜在合作方合作，点对点交换数据→私有数据公开发布，回馈公共平台的良性发展模式。

5 实践与思考

我们按照微生物组大数据中心的设计理念，开展了前期工作，在组学数据百科全书 NODE（http://www.biosino.org/node/）的支持下，建立了微生物组数据专区（http://www.biosino.org/microbiome）和微生物组分析平台（http://www.biosino.org/microap）。微生物组数据专区主要提供微生物组的公共组学数据的浏览、查询与发布，并选择有代表性的数据作为参比数据，为微生物组分析平台提供支撑。微生物组分析平台依赖于数据专区中的参考数据，提供元基因组功能分析和生态菌群的结构分析，后续将直接支持用户多种途径的全基因组测序数据分析，探索私有数据在公共平台上的保护与利用模式。

当然，微生物组大数据中心不是微生物组研究工作的全部，也不是微生物学研究的全部，甚至不是微生物组研究支撑平台的全部。因此，在建设微生物组大数据中心的同时，要特别强调数据中心与各种“实体库”和创新技术的互动。在微生物系统组方面，要特别注意与微生物菌种库的结合、协同发展。在微生物组方面，要与微生物组样本库（包括菌群库和 DNA 库等样本）结合、协同发展。而在与合成生物学元件库建设结合过程中，in silico、in vitro 和 in vivo 技术的结合，以工程学理念带来的设计-合成-测试概念与数据的引入，都将为我们跨入电子细胞 E^2（electoral/engineered）-cell 和 X/M-（extensive/multiple）-cell 时代奠定基础。

总之，符合上述特征的微生物组大数据中心，不仅能保证大型科研计划的顺利实施，还将长期为基础微生物、人口健康、环境等领域的广大科研人员提供丰富的数据集和不断更新发展的工具平台，将集聚他们的智慧与工作结晶，有力推动整个微生物学的发展以及微生物知识与技术的广泛深入的应用。

致谢 本文撰写过程中，得到了中科院深圳先进技术院合成生物学工程研究中心马迎飞研究员、中科院巴斯德所郝沛研究员的大力支持，特此致谢。

参考文献

1 McGuire A L, Colgrove J, Whitney S N, et al. Ethical, legal, and social considerations in conducting the Human Microbiome Project. Genome Res, 2008, 18(12): 1861-1864.

2 Integrative HMPRNC. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe, 2014, 16(3): 276-289.

3 Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010, 464(7285): 59-65.

4 Cardona S, Eck A, Cassellas M, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. BMC Microbiol, 2012, 12: 158.

5 Mitchell A, Bucchini F, Cochrane G, et al. EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res, 2016, 44(D1): D595-603.

6 Sunagawa S, Coelho LP, Chaffron S, et al. Ocean plankton. structure and function of the global ocean microbiome. Science, 2015, 348(6237): 1261359.

7 Mukherjee S, Stamatis D, Bertsch J, et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res, 2017, 45(D1): D446-D456.

8 Chen I A, Markowitz V M, Chu K, et al. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res, 2017, 45(D1): D507-D516.

9 Meyer F, Paarmann D, D’Souza M, et al. The metagenomics

RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 2008, 9: 386.

10 Felippes F F, Wang J W, Weigel D. MIGS: miRNA-induced gene silencing. Plant J, 2012, 70(3): 541-547.

11 Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol, 2008, 26(5): 541-547.

12 Ten Hoopen P, Pesant S, Kottmann R, et al. Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. Stand Genomic Sci, 2015, 10: 20.

13 Vandamme P, Pot B, Gillis M, et al. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev, 1996, 60(2): 407-438.

14 Ramasamy D, Mishra A K, Lagier J C, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol, 2014, 64(Pt 2): 384-391.

15 Soucy S M, Huang J, Gogarten J P. Horizontal gene transfer: building the web of life. Nat Rev Genet, 2015, 16(8): 472-482.

16 Vandamme P, Peeters C. Time to revisit polyphasic taxonomy. Antonie Van Leeuwenhoek, 2014, 106(1): 57-65.

17 Cole J R, Wang Q, Fish J A, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res, 2014, 42(Database issue): D633-642.

18 McDonald D, Price M N, Goodrich J, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012, 6(3): 610-618.

19 Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res, 2007, 35(21): 7188-7196.

20 Yarza P, Richter M, Peplies J, et al. The all-species living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol, 2008, 31(4): 241-250.

21 Mou X, Sun S, Edwards R A, et al. Bacterial carbon processing by generalist species in the coastal ocean. Nature, 2008, 451(7179): 708-711.

22 Warnecke F, Luginbühl P, Ivanova N, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature, 2007, 450(7169): 560-565.

23 Schloss P D, Westcott S L, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol, 2009, 75(23): 7537-7541.

24 Zhou Q, Su X, Jing G, et al. Meta-QC-Chain: comprehensive and fast quality control method for metagenomic data. Genomics Proteomics Bioinformatics, 2014, 12(1): 52-56.

25 Zhou Q, Su X, Wang A, et al. QC-Chain: fast and holistic quality control method for next-generation sequencing data. Plos One, 2013, 8(4): e60234.

26 Finotello F, Mastrorilli E, Di Camillo B. Measuring the diversity of the human microbiota with targeted next-generation sequencing. Brief Bioinform, 2016.

27 Huson D H, Auch A F, Qi J, et al. MEGAN analysis of metagenomic data. Genome Res, 2007, 17(3): 377-386.

28 Caporaso J G, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods, 2010, 7(5): 335-336.

29 Su X, Xu J, Ning K. Parallel-META: efficient metagenomic data analysis based on high-performance computation. BMC Systems Biology, 2012, 6(Suppl 1): S16.

30 Skinnider M A, Dejong C A, Rees P N, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic acids research, 2015, 43(20): 9645-9662.

31 Dejong C A, Chen G M, Li H, et al. Polyketide and

nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nature chemical biology, 2016, 12(12): 1007-1014.

32 Klementz D, Doring K, Lucas X, et al. StreptomeDB 2.0--an extended resource of natural products produced by streptomycetes. Nucleic acids research, 2016, 44(D1): D509-514.

33 Blin K, Medema M H, Kottmann R, et al. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Res, 2017, 45(D1): D555-D559.

34 Placzek S, Schomburg I, Chang A, et al. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res, 2017, 45(D1): D380-D388.

35 Morgat A, Lombardot T, Axelsen K B, et al. Updates in Rhea - an expert curated resource of biochemical reactions. Nucleic Acids Res, 2017, 45(D1): D415-D418.

36 Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res, 2017, 45(D1): D353-D361.

37 Hu Q N, Deng Z, Hu H, et al. RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. Bioinformatics, 2011, 27(17): 2465-2467.

38 Hatzimanikatis V, Li C, Ionita J A, et al. Exploring the diversity of complex metabolic networks. Bioinformatics, 2005, 21(8): 1603-1609.

39 Moriya Y, Shigemizu D, Hattori M, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res, 2010, 38(Web Server issue): W138-143.

40 Pitkanen E, Jouhten P, Rousu J. Inferring branching pathways in genome-scale metabolic networks. BMC Syst Biol, 2009, 3: 103.

41 Tu W, Zhang H, Liu J, et al. BioSynther: a customized biosynthetic potential explorer. Bioinformatics, 2016, 32(3): 472-473.

42 King Z A, Lloyd C J, Feist A M, et al. Next-generation genome-scale models for metabolic engineering. Curr Opin Biotechnol, 2015, 35: 23-29.

43 Chou C H, Chang W C, Chiu C M, et al. FMM: a web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res, 2009, 37(Web Server issue): W129-134.

44 Rahman S A, Advani P, Schunk R, et al. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics, 2005, 21(7): 1189-1193.

45 Kuwahara H, Alazmi M, Cui X, et al. MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind. Nucleic Acids Res, 2016, 44(W1): W217-225.

46 Thiele I, Palsson B O. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc, 2010, 5(1): 93-121.

47 Narihiro T, Sekiguchi Y. Microbial communities in anaerobic digestion processes for waste and wastewater treatment: a microbiological update. Current opinion in biotechnology, 2007, 18(3): 273-278.

48 Tong X, Xu H, Zou L, et al. High diversity of airborne fungi in the hospital environment as revealed by meta-sequencing-based microbiome analysis. Sci Rep, 2017, 7: 39606.

49 Tringe S G, Zhang T, Liu X, et al. The airborne metagenome in an indoor urban environment. Plos one, 2008, 3(4): e1862.

50 Gupta R, Beg Q K, Lorenz P. Bacterial alkaline proteases: molecular approaches and industrial applications. Appl Microbiol Biotechnol, 2002, 59(1): 15-32.

51 Cephas K D, Kim J, Mathai R A, et al. Comparative analysis of salivary bacterial microbiome diversity in edentulous infants and their mothers or primary care givers using pyrosequencing. Plos one, 2011, 6(8): e23503.

52 Yang F, Zeng X, Ning K, et al. Saliva microbiomes distinguish caries-active from healthy human populations. The ISME

Journal, 2012, 6(1): 1-10.

53 Jumpertz R, Le D S, Turnbaugh P J, et al. Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. The American Journal of Clinical nutrition, 2011, 94(1): 58-65.

54 Kong H H. Skin microbiome: genomics-based insights into the diversity and role of skin microbes. Trends Mol Med, 2011, 17(6): 320-328.

55 Vasudevan D, Richter H, Angenent L T. Upgrading dilute ethanol from syngas fermentation to n-caproate with reactor microbiomes. Bioresour Technol, 2014, 151: 378-382.

56 Shi W, Moon C D, Leahy S C, et al. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res, 2014, 24(9): 1517-1525.

Development of Comprehensive Microbiome Big Data Warehouse/Center for Long-term Scientific Impact

Zhang Guoqing[1] Ning Kang[2] Zhi Xiaoyang[3] Liu Wan[4] Xu Ping[5] Zhou Haokui[6] Hu Qiannan[1] Zhao Guoping[1]

（1 Bio-Med Big Data Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;

2 The School of Life Science and Technology, Huazhong University of Science & Technology, Wuhan 430074, China;

3 Yunnan Institute of Microbiology, Yunnan University, Kunming 650091, China;

4 Shanghai Center for Bioinformation Technology, Shanghai 201203, China;

5 Shanghai Information Center for Life Sciences, Chinese Academy of Sciences, Shanghai 200031, China;

6 Center for Synthetic Biology Engineering Research, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China）

Abstract It was the scientific concept and related technology of metagenomics that initiated the microbiome research. These microbiome research projects conducted globally have led to the acquisition huge amount of data and data sets of microbial genomes related to human health, animals, plants and environments. Consequently, various kinds of microbiome databases and analytical platforms are booming. However, besides the designed specific project-oriented status for some of the databases, most of the current microbiome data platforms merely focus on the development of reference data catalog and metagenome data sets, and mainly support the studies of"molecular ecology"aspect of microbiomes and/or the metagenome of a specific biotype. Thus, commonly expected applications in data integration-dependent mega-analysis, genomic information-based microbial taxonomy or comprehensive functional bioparts mining are largely hindered by lacking of proper data resources or sophisticated bioinformaticians capable of handling the complicated tasks.In this review, we introduce the concept of *Microbiophylome*, which is the sum of all microbes and member organisms of all kinds of microbiota with their genetic and multiple life-omics information as well as their related biological structural/functional information. Comparing to the conventional *Microbiome*, which is the sum of all member microbes of various microbiota in a special ecological biotype with their genetic, mainly metagenome information and related biological function, *Microbiophylome* emphasizes the total information of every individual taxon of the whole microbial world. In other words, with respect to microbiology as an academic discipline, *Microbiophylome* is concerned more about the α-phase (taxonomy) and β-phase (phylogeny) of microbial biology while Microbiome is concerned more about the γ-phase (ecology), employing the knowledge of α- and β-phases. With the integration of the concepts of *Microbiome* and *Microbiophylome*, we suggest to establish a comprehensive microbiome data warehouse as a hub to integrate the data of microbial taxonomy, evolution and ecology as well as their related omics research. *Via* further

integration of the data of basic research in life science and systems and synthetic biology, this data warehouse will support the development of comprehensive and QA/QC controlled reference databases, high quality standards-guided assembly and annotation and state of the art tools for data integration, searching, shared analysis and deep mining to facilitate future academic research and biotechnology R&D activities in microbiology and related fields. In addition, providing high-quality data standard and data SOPs for safe data integration and sharing, this data warehouse will be attractive for further systematic collection of meta-data of large-scale international projects. We have started this effort aiming at the eventual establishment of a microbiome big data center with complete and integrative data storage, safe and efficiency-guaranteed data management as well as comprehensive and user-friendly data service functions.

Keywords microbiome, microbiophylome, classification, ecology, synthetic biology

张国庆 中科院上海生命科学院生物医学大数据中心副主任，研究员。主要研究领域包括：生物信息学数据库与知识库。长期致力于精准医学、大型人群队列、个性化药物研发、微生物组与合成生物学等领域的组学数据、文献数据和临床数据的整合与挖掘。E-mail: gqzhang@picb.ac.cn

Zhang Guoqing Vice director and principal investigator of Bio-Med Big Data Center of Shanghai Institutes for Biological Sciences of Chinese Academy of Sciences. Zhang's main research interest is bioinformatics database and knowledge base, focusing on the integration and mining of omics data, literature data and clinical data in the fields, such as precision medicine, large population cohort, the development of personalized drug, microbiome and synthetic biology etc. E-mail: gqzhang@picb.ac.cn

赵国屏 男，中科院院士，中科院上海生命科学院生物医学大数据中心首席科学家，植物生理生态所研究员，国家人类基因组南方研究中心执行主任，兼任中国微生物学会和生物工程学会理事会顾问。研究微生物代谢调控以及酶的结构功能关系与反应机理，开发相应的微生物和蛋白质工程生物技术。E-mail:gpzhao@sibs.ac.cn

Zhao Guoping Male, Academician of Chinese Academy of Sciences, chief scientist of Bio-Med Big Data Center of Shanghai Institutes for Biological Sciences of Chinese Academy of Sciences,professor of Institute of Plant Physiology and Ecology, executive director of the Chinese National Human Genome Center at Shanghai (CHGCS).Zhao is also counsellor to the Board of Chinese Society for Microbiology and Chinese Society of Biotechnology, and Shanghai Society for Microbiology. Zhao has been working on the structure function relationship and reaction mechanisms of microbial enzymes. Based on these studies, he is also interested in developing microbial and/or protein engineering technology for industrial application of these enzymes. E-mail:gpzhao@sibs.ac.cn